Clustering for Opportunistic Communication

Jay Budzik, Shannon Bradshaw, Xiaobin Fu, and Kristian J. Hammond

Department of Computer Science, Northwestern University
1890 Maple Ave.
Evanston, IL 60201 USA
+1 847 491 3500

{budzik, bradshaw, fu, hammond}@infolab.northwestern.edu

7/16/02

ABSTRACT

We describe ongoing work on I2I, a system aimed at fostering opportunistic communication among users viewing or manipulating content on the Web and in productivity applications. Unlike previous work in which the URLs of Web resources are used to group users visiting the same resource, we present a more general framework for clustering work contexts to group users together that accounts for dynamic content and distributional properties of Web accesses which can limit the utility URL based systems. In addition, we describe a method for scaffolding asynchronous communication in the context of an ongoing task that takes into account the ephemeral nature of the location of content on the Web. The techniques we describe also nicely cover local files in progress, in addition to publicly available Web content. We present the results of several evaluations that indicate systems that use the techniques we employ may be more useful than systems that are strictly URL based.

Categories and Subject Descriptors

H.5.3 [**Information Interfaces and Presentation**]: Group and Organization Interfaces – *asynchronous interaction, collaborative computing, computer-supported-cooperative work, evaluation/methodology, synchronous interaction, theory and models, web-based interaction.* H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *clustering.* H.3.5 [**Information Storage and Retrieval**]: Systems and Software – *current awareness systems, user profiles and alert services.* I.2.11 [**Artificial Intelligence**]: Distributed Artificial Intelligence – *intelligent agents.*

General Terms

Design, Experimentation, Human Factors

Keywords

Collaboration, clustering, awareness, agents, opportunistic communication, context, critical mass.

1. INTRODUCTION

Recent research has focused on making the activity of visitors to Web sites more visible (e.g., [6]). The motivation behind this work is that people browsing the same location on the Web could find it useful or entertaining to talk with each other. Several commercial ventures have developed robust implementations of systems that allow users who are visiting the same Web resource (as represented by its URL) to talk with each other using instant messaging or even videoconferencing (e.g., [8]). The access to communities these systems enable can provide a useful starting point for finding answers to questions regarding content, sharing ideas and comments, and meeting people with similar interests. The goal of this work is to support interactions among people that are opportunistic and are often based on a establishing a shared context for the interaction. In these systems, a shared context is established by visiting the same location on the Web.

Analyses of Web access logs have given us a better understanding of the distributional characteristics of access frequencies [1, 5, 11]. These studies show that just a few sites are visited very frequently, whereas the vast majority of sites are visited quite infrequently. Moreover, it is clear that many locations on the Web actually contain the same, or very similar content. For example, many news sites carry verbatim copies of the same story. In addition, dynamic content (generated by form submissions and cookies, for example) violates the assumption of a 1 to 1 relationship between a URL and the content of a document.

These facts have led us to question the overall utility of URL based approaches to building context-based communities. If every browser included a URL based collaboration system, current models of Web access behavior would predict that such systems would tend to return unmanageable numbers of users to talk to, or no one at all. In addition, users might be erroneously grouped together because of dynamic content. In light of this analysis, we propose an alternative to viewing the Web as a set of isolated locations represented by their unique URL. Specifically, we propose using richer representations of a user's context (of which the URL they are visiting might be included) coupled with methods of computing similarity among them. We argue context similarity is more flexible and can more easily accommodate both sparsely populated areas on the Web, as well as the crowded ones. We describe an instantiation of this framework in a system we have developed called I2I.

I2I automatically tracks the work contexts of distributed users as represented by the content of the documents they manipulate using standard productivity applications as well as Web browsers (we adopt a similar context tracking approach in [3]). The system clusters the documents they use based on their content, grouping related documents into a conceptual neighborhood, allowing users to:

1. Establish synchronous communication with others who are manipulating related documents.
2. Initiate conversations asynchronously through a facility we call calling cards.

WWW 2002, May 7-11, 2002, Honolulu, Hawaii, USA.
ACM 1-58113-449-5/02/0005.

3. Browse related information items automatically recommended by the system.
4. Join or start public chat rooms associated with the content area in which they are situated.

I2I attempts to manage the early stages of initiating informal collaboration by providing its users with opportunities to become aware of the activities of others that share common interests, as represented by the documents they interact with. I2I attempts to build communities of common interest on the fly, allowing users engaged in traditionally solitary activities to discover common goals and collaborate with each other, while reducing the overhead of orchestrating collaboration.

Our aim is for I2I to make opportunities for informal collaboration more obvious and more pervasive by reducing the amount of work necessary to become aware of them. The motivation is the possibility that by increasing awareness of common work contexts, users will leverage each other's knowledge and experiences more frequently, which could allow them to be more productive. The problem is finding a balance between the benefits of having access to a large, diverse body of people and the level of effort necessary to find someone helpful. On the one hand, access to large numbers of people means someone relevant is probably out there. Current electronic communication systems give us that. Unfortunately it could take a significant amount of time to find the right person through traditional methods, a cost that often far outweighs the benefits a user might expect. I2I is designed to automate part of this process by noticing opportunities for collaboration based on the work people do in everyday applications. It provides a first cut at helping users discovering potential collaborators by giving users opportunities to become aware of others who are manipulating similar documents. Combined with standard communication tools, our goal is for a system like I2I to routinely transform traditionally solitary activities into collaborative ones by providing its users with frictionless access to potentially relevant others.

1.1 Example of Use

It is instructive to consider an example of how the system could be used in order to better understand the utility of the techniques we describe. For example, say Mary and Joe are both high school students. Joe is writing a term paper for his ecology class about the environmental impact of pesticides. Joe is from a small town in rural New York, and has first-hand knowledge of this issue from the summers he spent working on a farm. Mary is writing a letter to her congressperson about an upcoming bill that would provide tax incentives to farmers who adopt more environmentally friendly practices. Mary lives in New York and is the president of her school's Earth First chapter, which, among other things, promotes consumer awareness of the benefits of pesticide-free farming. Mary and Joe are both using I2I, and as such, the system notices they are writing about similar subjects and displays their screen names in a window associated with their current document. Mary and Joe can now contact each other through I2I using text messaging or videoconferencing. The system provides them with an awareness of their shared work contexts and interests, which serves as common ground for the conversation they start about their writing.

1.2 Relation to Previous Work

Tools that allow users to collaborate around common electronic artifacts have been studied extensively, although much recent work has focused on collaboration around documents. Anchored Conversations [4], for example, allow collaborators to easily distribute shared documents and situate conversations within the context of specific places in a shared document. Ensuring collaborators share the same artifact makes collaborative activities that depend strongly on artifacts (such as collaborative writing) easier.

Systems like this are aimed at supporting collaboration among users who already know each other and have a prior goal to collaborate. Our work intends to provide opportunities for users who may not know each other to collaborate informally by making opportunities for collaboration visible, and by automating the early stages of establishing collaboration (e.g., knowing who to talk to, and how to talk with them). In this way, the work we describe here is similar to Jung and Lee's work on ePlace, a system aimed at providing a rich environment that supports mutual awareness among visitors of e-commerce sites [12]. This work has resulted in particularly clever visual representations of Web sites, on top of which information is overlaid, indicating the presence of visitors. Our work differs from theirs in that it strives to extend beyond using the (somewhat arbitrary) structural organization of the Web and Web sites to establish awareness. Instead, I2I's notion of location is based on the content of the information objects users manipulate. This allows the system to notice similar work contexts that are not explicitly associated by hyperlinks (or even published on the public Web). In addition, the framework we describe is flexible enough to allow additional contextual information to be added to the computation of a neighborhood.

Other systems have examined the role of distributed, public artifacts (e.g., Web pages and online virtual environments) as shared contexts, allowing users who are manipulating or viewing the same object from distributed locations to communicate with each other, usually using text-based chat (e.g., [6, 8, 13]). This vein of work is the closest to the work we describe here. The main difference is that the above systems require objects tagged by unique identifiers (in the case of the Web, the page's URL), and also that users manipulate the same object at the same time in order to collaborate. These requirements limit the opportunities for collaboration the system can make available, due to the sheer size of the Web and typical patterns of access. Web accesses have been shown to follow a Zipf distribution, which means it is unlikely two users will be on the same page at the same time except at the most popular sites [1, 11]. The most popular sites are typically portals to other content like search engines and Web directories, and hence are not ideal as a shared context for collaboration. Likewise, it is also clear that similar content can be found on many different URLs, and that documents with the same URL may contain different content (due to personalization features and form submissions). As such, systems that use document identity as their only basis for introducing people are limited in the number and quality of opportunities for collaboration they can notice. In addition, approaches that rely on public artifacts like Web pages exclude unpublished electronic documents from consideration. Together these issues limit the utility of such systems.

I2I, first reported in [2], attempts to overcome these problems by clustering user contexts. User contexts are weighted vectors of features, which under the current system are comprised of terms derived from the textual content of the documents users manipulate. In effect, I2I builds a separate conceptual space,

organized by the content of its user's documents, and then situates users and other information items in this space. In addition, I2I provides facilities for asynchronous communication, allowing users to notice opportunities for collaboration across time.

I2I is also related to matchmaking systems (e.g., [7, 10]), which introduce users with common interests to each other with the goal of building online communities and fostering community awareness. Work on matchmaking systems has generally focused on introducing users based on long-term interests represented in a user profile. In the Yenta system [7], for example, users submit a collection of documents to their personal agent, which builds a profile from those documents and executes a kind of distributed hill-climbing algorithm to match profiles using techniques from information retrieval. When a sufficiently similar profile is found, the agent arranges to introduce the two users. The matching algorithms used in such systems are typically designed to work asynchronously.

In contrast, our work on I2I focuses on introducing users based on their immediate (and perhaps short-lived) interests that arise from the tasks they in which they are currently engaged. Instead of requiring the users define a profile for themselves using documents or keywords, I2I automatically builds a lexical representation of the user's current activity (as represented by the document the user is manipulating) and uses this representation to determine what the user can see, as they are working.

2. CONTEXT SIMILARITY AS A FRAMEWORK FOR INTRODUCTION

I2I uses the work contexts of its users to determine whether or not it should make an introduction. Contexts are represented as vectors of features in a feature space $F = (f_1, f_2, \ldots, f_k)$, where each f_i is a number representing the weight or influence of the i^{th} feature. Each context has an owner, the user who is performing the activity the context represents. Given two contexts, A and B both vectors in F, we can define a function $d : F \times F \rightarrow [0, 1]$, which determines the distance or similarity between A and B. Given the user's current context A, d allows us to compute an ordering of the other contexts the system has captured. This ordering allows us both implement a policy in which we present users with contexts similar above a threshold θ, if we are concerned with relevance, or the users corresponding to the top n contexts if we are concerned with making the system maintain a constant number of users, or some combination of both.

I2I currently uses word stems as the feature space F. A context is defined by the word stems in the document the user is manipulating, weighted using the *tfidf* heuristic [17], which values words that are frequent in the current document, but rare across the collection of documents. Word stems are computed using the Porter suffix-stripping algorithm [15], which transforms multiple word forms (e.g., running, runs) into a single base form (e.g., run). The vectors representing each document correspond to points in a high-dimensional space; the number of unique word stems in the active documents determines the dimensionality of the space; the weighting heuristic determines the position of a vector in the space. The distance d between two contexts is defined as the cosine of the angle formed by the two vectors that represent them. Essentially, the cosine is a weighted function of the features the vectors have in common. The system uses a threshold policy to present the top 10 users corresponding to contexts similar above $\theta = 0.65$, a value determined empirically to best balance the trade-off between relevance, and the likelihood of seeing someone else (as we go on to describe in following sections).

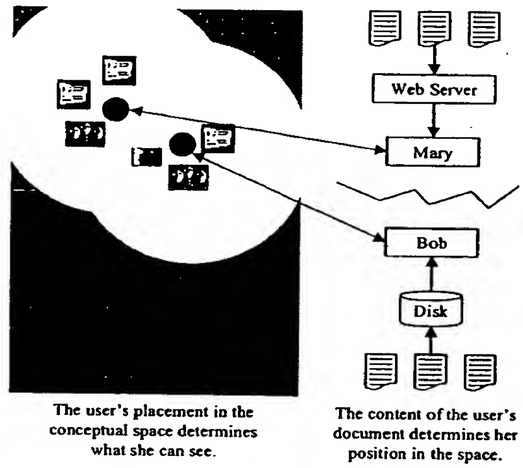

Figure 1. I2I clusters users together based on the documents they are manipulating.

Note that under this framework if we define the feature space F to be the set of all legal URLs, let each user's vector contain a 1 in the position of the URL they are visiting and zeroes elsewhere, use a threshold policy in which we present all users associated with contexts similar above $\theta = 0.5$, and let

$$d(A, B) = \begin{cases} 0, \text{if } A = B \\ 1, \text{if } A \neq B \end{cases},$$

then we have defined the introduction policy of several previous systems [6, 13].

Different definitions of F and d might take into account other elements of context, such as the user's long-term interests represented by her browsing history, membership in groups related by a semantic network, her level of activity, or whether or not two users have used the system to talk before. In addition F and d can be defined to apply to different kinds of documents altogether (e.g. figures), or leverage other metrics like hyper-link distance (number of links one must traverse between Web-based documents). One of the most crucial aspects of designing a system like this is determining a representation of context that provides adequate performance in the context of the goals of its users. Future work is aimed at developing richer representations of the user's context in order to provide users with even more useful opportunities for collaboration. In addition, we look forward to the promise of the semantic Web. Richer representations of content and their relationships to each other will allow us to develop better representations of user context.

It is important to note that any object can be associated with a representation of a context in this framework. The simplest example is the context's owner (the user who is browsing or writing the document represented by the context vector). But chat rooms, newsgroups, graphics, and collections of other documents all can be explicitly associated with a context. The I2I system makes use of this feature of reified context representations in order to support the various communication modalities described in the following section.

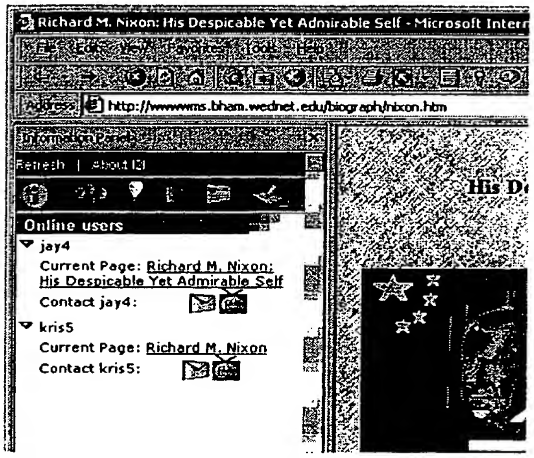

Figure 2. The I2I user interface embedded in Microsoft Internet Explorer.

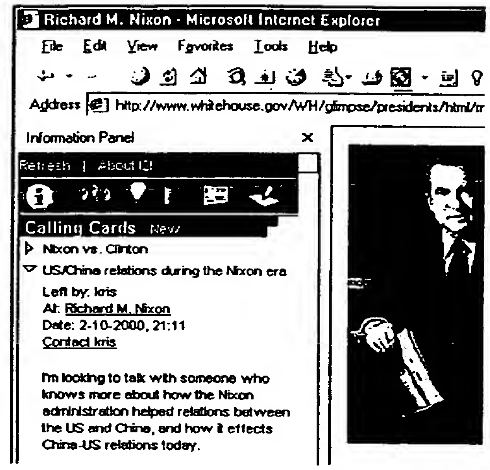

Figure 3. Details of I2I's Calling Card Interface

3. SYSTEM ARCHITECTURE

I2I integrates with applications through their APIs and the operating system's inter-process communication facilities. Each application has a corresponding application adapter (as in [3]), which is responsible for communicating user actions and document content to a broker, located on a central server. The broker is responsible for persistent information such as the user's profile (e.g., their name, password, etc.), as well as ephemeral information, such as how to contact their machine, and the representation of the document they are currently manipulating.

Each application adapter is responsible for sending the broker a message when the document has changed in an attached application (e.g., the document is edited significantly, or the user opens or navigates to a new one), as in [3]. The message the adapter sends to the broker (located on a central server or server cluster) contains the text of the active document, its location (URL), and the user ID of the I2I user.

The broker computes a vector representing the user's current context (the document she is manipulating) as described above. Each user is associated with a vector (or several, if they have multiple documents open). Associated with each vector is the title of the document it represents, the URL (if the document is a Web page), a list of users manipulating that document, chat rooms started from that document, as well as a list of calling cards associated with the document.

The broker computes a pair-wise similarity matrix for documents that are currently in use, which it maintains in memory for fast updates. The chance of two people reading exactly the same document at the same time may be slim. By grouping conceptually similar documents together, I2I makes it more likely that people will see each other and start a conversation. It also allows unpublished documents (e.g., a paper in progress) to serve as the entry point into the system.

Secondary objects can also be associated with a document in the space I2I has built. The simplest of these objects is people: users who are viewing a particular document are associated with that document's point in the space. Currently, I2I also indexes chat rooms and calling cards (a facility for asynchronous communication) in the same way. Users who access a document, then, can see the items associated with it and other documents close to it in the context space (see Figure 1).

4. USER INTERFACE

I2I tracks a user's current task context (represented by the document they are manipulating) so it can provide potentially useful resources to users in the context of a specific editing or browsing session. It embeds an interface for displaying this information directly into applications, where it is supported (see Figure 2) to allow the user to easily correspond the information I2I displays with the document it is associated with. In other cases, information is displayed in an associated window that can be "hooked" on to the main window of the application, to maintain visual correspondence. This allows users to easily keep track of their activity in several conceptual spaces at the same time. The primary interface for I2I is written in DHTML.

Details of the embedded interface are shown in Figures 2 and 3. Information is grouped into tabs and includes (from left to right):

1. System activity. Users can see how many I2I users are online both in and outside of the conceptual space defined by their document. Other activity information includes how many people are chatting, and how many related pages other I2I users are reading.
2. Who is online. Users can see the login names of the people reading or writing related documents, and pointers to the documents they are viewing, if they are available on the Web (see Figure 2). Users can contact each other directly via instant messaging, double-blind email, or by using videoconferencing software, depending on the software and hardware available on their machine.
3. Related documents. I2I displays related pages from other Web sites people are currently browsing. In addition, I2I displays recommendations generated by the Watson system [3]. The Watson system recommends related documents by automatically querying online information repositories.
4. Active chat information. I2I displays a list of chat topics created by users within the conceptual space defined by the

current document. Users can also chat in a default room associated with this region in the content space.

5. Calling cards. I2I displays a list of calling cards that other users have left in the past while viewing the current or related documents (see Figure 3). A calling card is a note associated with a region of the content space I2I builds that indicates a user would like to talk about a particular topic.

I2I's interface is designed to allow its users to quickly become aware of people working on similar content. Users can contact others who are working on similar content at the same time they are by browsing the "Who is Online" tab. In addition, the calling card functionality (described in more detail below) allows users to contact each other across time.

4.1 Calling Cards

Users can leave calling cards associated with the content area represented by their document in order to indicate they would like to discuss specific aspects of that topic with other users. Figure 3 shows calling cards associated with a page about Richard Nixon.

Leaving a calling card allows users to make their goal to discuss a topic visible to other users who also view documents in that topic area. If a user is eager to open a discussion channel with somebody else, but no one is available or has shown interest, the user can leave a message to invite people to talk at a later date. After leaving a calling card, the user can continue to work, or even destroy the original document the calling card was associated with. A calling card is indexed by the context vector that represents the document at which it was created. This means access to the document is not necessary for other users to see its associated calling cards when they are browsing or writing in related areas.

For example, one user could leave a calling card at the document in Figure 2, which discusses Nixon and his presidency. That document could then be taken off of the Web. At this point, other users would still be able to see the calling card when they accessed other documents about Nixon, for example, the page in Figure 3.

This kind of contextual indexing also nicely accommodates documents that have frequently-updated content (like the front page of a news site), because even though the content of a page might change, the system associates the calling card with the original context in which it was created. This ensures that the calling cards retrieved are actually relevant in the context of the document being viewed. This approach is similar to the independently developed intra-document linking technique reported and evaluated in [14].

When another user sees the calling card, she can find out whether the owner of the calling card is online or not. If the owner is online, she simply drops a line to the user to say that she is interested in discussing the topic. If the owner is not online, she can find out if the owner has a public email address and send an email to the owner (if the owner has specified others can contact her via email).

A calling card persists for a time period specified when it is created (currently the system imposes a limit of 30 days). When a calling card expires, the owner is notified via the global interface. The user can then choose to delete the card, or extend the time period in which it is available.

4.2 Managing Privacy

Users may, at times, be uncomfortable with having a system track what they write or view. I2I allows people to manage the privacy of their work by being highly visible when it is on (see Figures 2 and 3), and by allowing users to shut it off at any time (using the close button in the interface). I2I also does not expose the details of offline (non-Web) content to any third party. In addition, it does not expose a user's email address or online identity directly. Instead email is sent through a mediating server that automatically makes message sender and recipient anonymous. This allows users to disclose their real email addresses at their own discretion.

Pertinent future work includes allowing users to adjust the extent to which their identity is revealed by developing online trust relationships with other I2I users. The idea is that users who have no prior affiliation can choose to reveal elements of their user profile (for lack of a better term) to each other, even though the full profile could be used to contribute to the similarity score the system computes. In general, we anticipate mitigating privacy concerns will be a significant issue moving forward. As such, we are working on facilities that allow users to maintain awareness and control over their trust relationships with both the users they know and users they haven't met.

4.3 Global Interface

I2I has a global interface that allows the user to control whether or not she is available for conversation, as well as edit her profile and preferences. Calling cards are also managed using this interface (owners of calling cards can edit or delete them at any time).

5. EVALUATION AND ANALYSIS

I2I is still under development, and although it is a suitable demonstration system, it is not ready for deployment. Thus, in lieu of a field study, in which we could get a broader sense of the system's effectiveness in actual work contexts, we have instead evaluated the matching techniques I2I uses on real data collected from users. We collected about two days of browsing logs from 11 users and performed an offline analysis. The browser logs were collected via a plug-in to Internet Explorer that recorded the URL, the time of access, and the content of a document when it was loaded. The users were graduate students at Northwestern (either in the School of Education or in the Computer Science department), or friends of the graduate students who participated.

Some of the following evaluations are performed by sampling the original distribution of accesses to simulate the use of the system by varying numbers of users. Because the distribution we observed follows larger distributions in character, the simulated results we describe are likely to be predictive of the system's performance with larger numbers of users. The one caveat is that the distribution of the content we collected may not match the content distribution of the Web in general, because the subjects who produced this data were not chosen at random from the collection of all Web users. We are not prepared to argue that the interests of graduate students are representative of the interests of the more general population of potential I2I users. However, one deployment strategy we are considering for I2I is aimed at facilitating awareness of the activities in others strictly within organizations. In this case, we would expect this data would be fairly representative (in distribution and structure, if not in content).

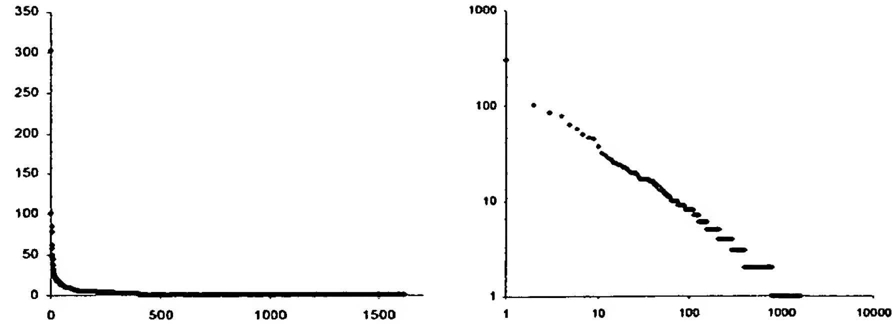

Figure 4. Distribution of access frequencies ordered by frequency. The graph to the right is the log-log plot of the same data.

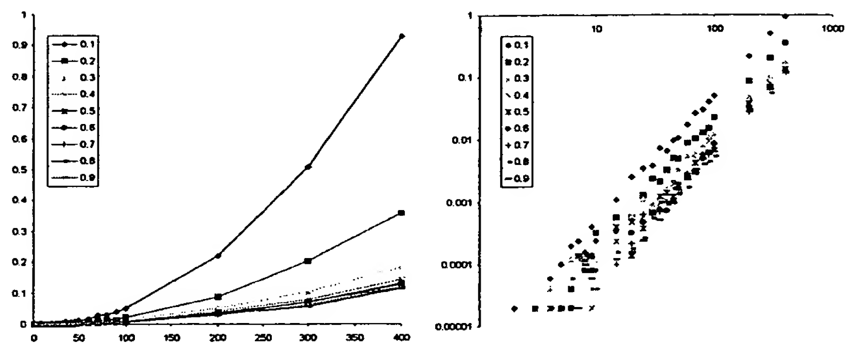

Figure 5. Left: number of simulated people (*x* axis) vs. likelihood an additional person will become visible if one is added (*y* axis) at 9 thresholds (series). Right: the same graph plotted on a log-log scale.

Note also that the content and frequency distributions of the data we gathered about users accessing documents on the Web is likely very different from their patterns of document access in word processors. In the future, we intend to extend these studies to include data gathered from users accessing and modifying documents in word processors.

That said, the analysis we performed attempted to achieve two goals. The first was to understand the relationship between the number of people using the system and the number of people they would see, on average, for varying levels of strictness in similarity. This analysis allowed us to gauge the number of users the system requires to begin providing contacts, essentially providing us with an idea of the "critical mass" requirements the system has (we use critical mass here as in [9]). The second goal was to evaluate whether a system using the techniques we describe above would make appropriate associations from the perspective of a potential user. Our results point clearly to a tradeoff between making the system present other people to the user and ensuring their current work contexts are sufficiently similar.

5.1 Summary of the Data

During the period we collected data, there were 1612 pages with unique URLs collected. This is only a lower bound on the number of unique pages viewed, because the same URL can contain different content (due to a form submission, for example). These pages were accessed a total of 5039 times.

As previous work (e.g., [11]) suggests, Web access data follow a Zipf distribution [18]. That is, if the frequency a page with frequency rank i is f_i, where the frequency rank i is the index of the i^{th} element in the sequence of documents accessed by descending frequency, then the Zipf's Law states $f_i \propto i^{\beta}$, where β close to -1. The data we gathered follow this distribution, with $\beta = -0.79$ ($r^2 = 0.96$). The linear and log-log plots of this data displayed in Figure 4 are typical.

5.2 Critical Mass Analysis

The most important consequence of the fact that document accesses follow a Zipf distribution is that a large number of documents are accessed relatively infrequently for any given period of time. For the data we collected, 1322 pages were accessed below the mean frequency of 3.12 (that's over 80%). About 50% of the documents were only accessed once. The data we collected supports our hypothesis that systems that provide users with information on who is browsing the *same* page would suffer serious "critical mass" problems. That is, at a given time, very few people will be present on anything but the most popular sites, leading to situations in which the system displays unmanageable numbers of people, or no one at all. The goal of the techniques we use for grouping people in I2I is to strike a more workable balance.

It is also important to note that the most popular pages in our set contain little lexical content (one of the top documents was the front page of a search engine with a markedly sparse interface). This follows results on a larger data set, reported in [5], which suggest smaller documents are accessed more frequently. More lexical content tends to improve the quality of match because words disambiguate each other's meaning (see [16] for an early

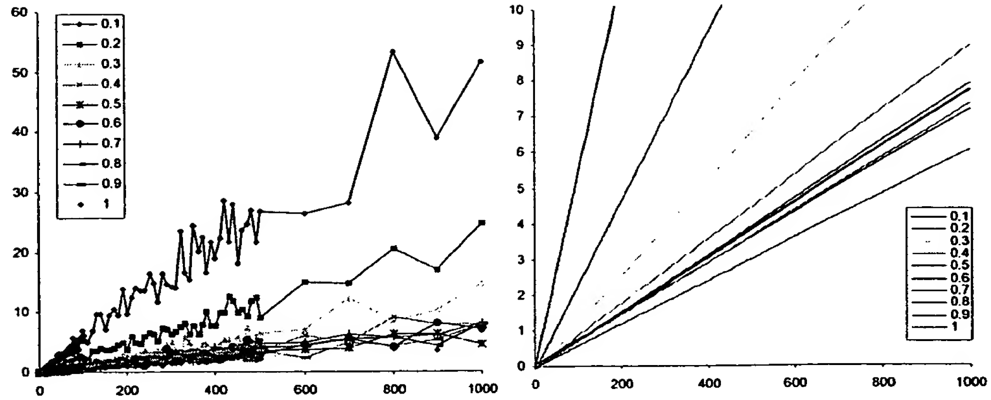

Figure 6. Number of simulated people vs. average number of people they would see from a particular page. Left: actual data; Right: best fit lines (least squares regression).

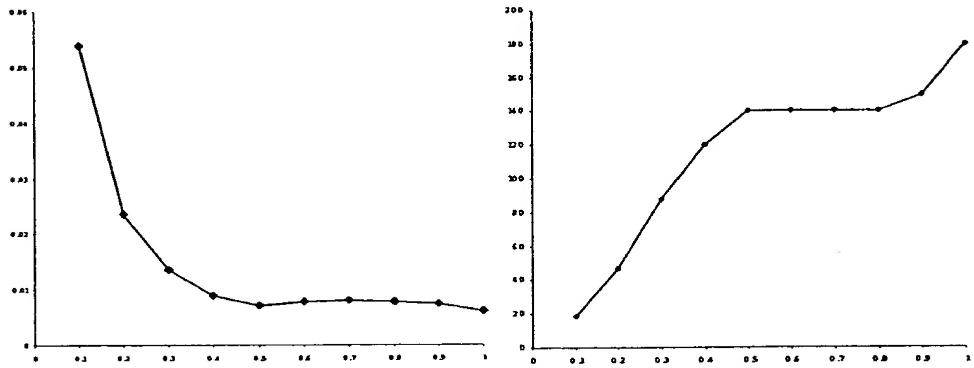

Figure 7. Left: Slope of the regression lines in Figure 12 vs. threshold. Right: number of people that must be using the system to see one other person on average, vs. threshold (interception of the regression line with $y = 1$).

technique that exploits this). Fortunately, the data suggests the pages with the least amount of lexical content happen to be ones at which multiple users are most likely to be accessing at the same time. This means that even though the similarity metrics we use rely on lexical content, those documents with the least amount of lexical content are those most frequently visited and hence will be more likely to have visitors at the same time. In addition, the most popular sites, which tend to be search engines and directories, may not typically provide a very good shared context for interactions among users because visitors to such sites often have very different goals.

The first critical mass evaluation we performed was aimed at determining the relationship between the number of users in the system and the possibility they would see someone new if they came online. Our hope was that this would give us a general sense of how the system would perform with particular numbers of users. We expected that as the number of users increases, their coverage of the space would also increase (that is, there would be more pages for which a new visitor would see at least one other user).

To accomplish this, we produced iteratively greater random samples drawn from the original distribution of page accesses we collected, in order to simulate varying numbers of users. Then we computed the percentage of accesses that would be visible from the vantage point of each document in this set, given one of several similarity thresholds (again, drawing the next access from the original distribution). We repeated this 10 times for each sample of "users" and took the average. In addition, this same analysis was repeated for 10 thresholds. The results of this analysis are displayed in Figure 5. As expected, the analysis shows that looser thresholds cover more of the document space, and that as the number of users approaches the number of documents the space is also covered more. Figure 5 also shows the log-log plot of the same data which can be fit using least-squares regression with average $r^2 = 0.98$. The slopes are close to 0.5 for each threshold, and the x intercepts are increasing as the threshold increases. That is, as the threshold increases, the number of users that must be logged into the system before at least two users will see each other (the x intercept) also increases.

We also evaluated the relationship between the number of users using the system and the average number of people they could see. To do this, we performed the same random sampling from the original access distribution for varying numbers of people. We then computed the average number of people visible to each of these users given a fixed similarity threshold. We repeated this 100 times for each sample of "users" and took the average. We

performed this analysis on 10 thresholds, for numbers of simulated users. The results of this analysis are displayed in Figure 6. Figure 6 (left) shows the raw data for various thresholds. Perhaps more instructive are the regression lines displayed in Figure 6 (right) (the average r^2 is 0.94 for these fits). There are several interesting things about Figure 6 (right):

1. It shows that various clustering thresholds cause the number of people seen to diverge at different rates—faster for lower thresholds and slowest for the tightest ones. This is displayed graphically in Figure 7, which plots the slope of the regression line vs. the threshold.
2. It shows that clustering documents improves the chance of a user seeing someone, even at the strictest thresholds.
3. It shows the number of active users needed to start reliably seeing other people from a particular document at a particular threshold. For loose similarity thresholds, this number is low. For higher thresholds, the number of active users required increases. This is displayed graphically in Figure 7 (right), which plots the values at which the regression line for a threshold crosses the line $y = 1$. This information will be valuable in the future, as we start to work on building newer versions of the system that can automatically adjust thresholds.

5.3 Appropriateness Analysis

In order to evaluate the appropriateness of the associations the system made, we had a volunteer uninvolved with this project evaluate the associations by hand. For each of 9 thresholds, we picked 10 random documents (we will refer to these 90 documents as the *source* documents). For each of the source documents, we randomly selected 10 documents that were similar above the threshold (the *target* documents). The volunteer was instructed to compare each source and the target and count the number of inappropriate associations made by the system (he performed a total of 900 comparisons). The results of this experiment are displayed in Figure 8. The data shows that for thresholds greater than or equal to 0.4, the system forms appropriate associations between source and target documents a least 60% of the time. For thresholds greater than or equal to 0.7, the system forms appropriate associations over 90% of the time.

5.4 Discussion

Together, the above analyses provide us with a better understanding of the relationships between number of people, similarity threshold, and relevance. Given an understanding of these relationships, we can begin to design the system to address the strengths and weakness revealed by this analysis. The above experiments were immediately useful in determining we should set the similarity threshold at about 0.6 or 0.7 in order to balance the tradeoff between the desire to have the system allow users to see other people, and the desire for the associations made by the system to be of the highest relevance. In addition they provided an empirical justification for our original critique of URL based systems. Future work entails determining what level of accuracy users find useful; this will help us further tune the system, evaluate the clustering techniques in this context, and develop new user interface facilities aimed at allowing the user to form correct expectations about how the system will operate.

It is important to realize we make a number of assumptions in this design. The most major assumption is that the text of the user's current document corresponds to her current goals and interests at a useful level of abstraction. We recognize this is not always the case, for example, when a user clicks on a link by mistake. However, our working hypothesis is that these are special cases and that the interfaces we have provided allow users to avoid any confusion they may cause.

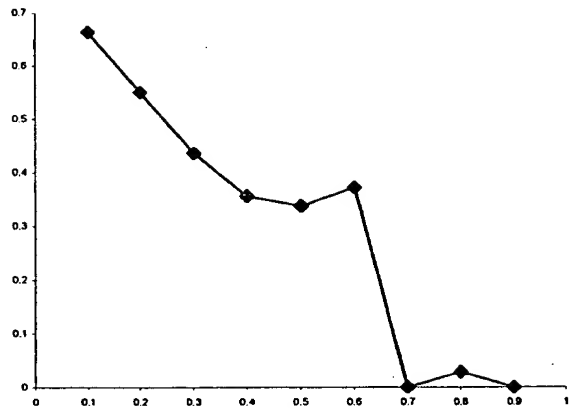

Figure 8. Threshold vs. percentage of inappropriate associations made by the system. As expected, as the threshold increases, the number of erroneous associations decreases.

The techniques we use for clustering documents have been shown to be effective (for example, similar techniques produce improvements in information retrieval [16]), but the above and other studies show that unintuitive associations can occur. It is also sometimes the case that the user's current document does not provide a very good window onto her goals (e.g., a single document can have multiple purposes). However, it is important to realize that the system does not require users to collaborate. It provides users with opportunities for collaboration by automatically recommending potential collaborators. In the end the user determines whether or not she takes the recommendation. Users can make their own decisions about whether to collaborate with each other based on their current needs and by inspecting the documents others are viewing (if they are available online), or by considering background information about the user (should users make such information available though the system). Improved interfaces for introduction are needed so that users who don't initially know each other can quickly determine whether or not spending the time to do so will be useful in the context of the tasks they are performing.

It is also important to note that the system provides the user with a representation of its current view of her work in the form of recommendations. If the system's recommendations are on-point, the user can be relatively certain the users recommended or the index terms for her calling card are also fairly appropriate. If the system displays off-point recommendations, then calling cards indexed in that context may be displayed in inappropriate places, and the users collected may be driven by unintuitive associations. Our studies show this will happen about 20% of the time [3]. The system's opt-in nature helps ameliorate the usability issues this causes. Likewise, examining the quality of the recommendations the system gives can also serve as a benchmark against which users can build accurate expectations. However, future iterations of I2I will most certainly expose more of the system's internal

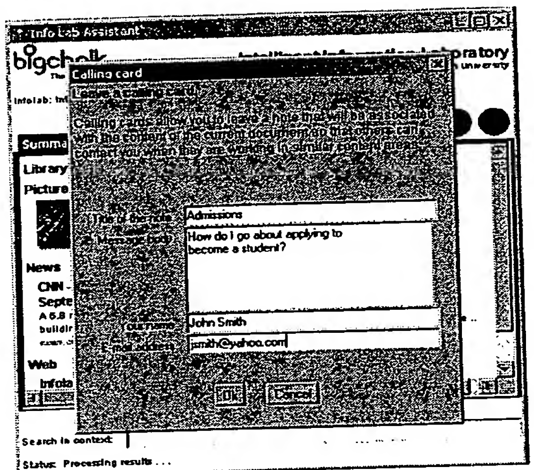

Figure 9. I2I's calling card functionality integrated with the Watson information access assistant.

representations of confidence to the user so that she can build better expectations about how the system will perform.

6. ONGOING AND FUTURE WORK

We have distributed I2I to several researchers in our department for limited use. For the most part, the feedback was positive. Users said they like the sense of being in a community and enjoy the kind of ready connectivity brought by I2I. We were encouraged by this initial test and look forward to a larger-scale deployment effort. This deployment effort is taking place in the context of extensions to the Watson information access assistant [3]. Of the features we implemented in I2I, perhaps the best received among potential users has been the Calling Card functionality, which allows users to explicitly register their interest in talking with someone about a particular content area. We have recently integrated this functionality within Watson and plan to release this functionality to pilot users in the near future (see Figure 9).

We are also working on improving the techniques we use to cluster work contexts. The evaluations uncovered several technical difficulties with the document similarity techniques we use that make them inadequate for handling some Web documents (e.g., URLs should be included as "terms" in a document so that pages with a single client-side image map can be coherently handled). In addition, we are investigating techniques for filtering potential collaborators by profiles built from their long-term history of interacting with documents (so that people with similar backgrounds are preferred).

Perhaps most compelling, however, is trying to understand what mix of people will benefit each other the most in the context of particular tasks. For example, a student stuck on a problem is likely to find another student who has finished that problem more helpful than someone who is also stuck. More generally, one aspect of a good collaboration is that it brings together people whose knowledge, skills, perspectives, and interests compliment each other in ways that are mutually beneficial. Giving the system better models of groups and individuals could allow it to automatically build this kind of complimentary collection of people. We see this as a particularly compelling direction for the future of this work.

7. CONCLUSION

The ubiquity of the Internet is changing the way people access information and the dynamics of how they interact with each other online. However, we can only take advantage of the resources available to us in this networked world if we are aware of them. I2I is aimed at facilitating an awareness of the resources available online to a user in the context of her current task. Our hope is that these kinds of awareness cues can help users by reducing the friction required to access resources instrumental to their task. I2I embeds communication facilities in the user's everyday applications so that users who share common work contexts can become aware of each other and communicate, even though they may have never met or discussed the interests they share. As we work to deploy this functionality in robust implementations, we are excited by its potential to positively change the way people work by allowing them to more easily leverage the resources available to them.

8. ACKNOWLEDGEMENTS

The authors would like to thank Andrei Scheinkman and Mason Warner for their work on an early prototype of this system. Peter Dinda provided an initial pointer to the Web access work. In addition this work has benefited from the comments of Larry Birnbaum, other members of the Intelligent Information Laboratory at Northwestern University, and anonymous reviewers.

9. REFERENCES

[1] Breslau, L., Cao, P., Fan, L., Phillips, G., and Shenker, S., "On the Implications of Zipf's Law for Web Caching," in *Proceedings of IEEE INFOCOM '99*, (New York, USA) 1999.

[2] Budzik, J., Fu, X., and Hammond, K., "Facilitating Opportunistic Communication by Monitoring User Activity in Everyday Applications," *CSCW 2000 Workshop on Awareness and the WWW*, (Pittsburgh, PA, USA), Available at http://www2.mic.atr.co.jp/dept2/awareness/ (Accessed Feb. 20, 2002).

[3] Budzik, J., and Hammond, K. J., "User Interactions with Everyday Applications as Context for Just-in-time Information Access," in *Proceedings of The 2000 International Conference on Intelligent User Interfaces*, (New Orleans, Louisiana, USA), ACM Press, 2000.

[4] Churchill, E. F., Trevor, J., Bly, S., Nelson, L., and Cubranic, D., "Anchored Conversations: Chatting in the Context of a Document," in *Proceedings of CHI 2000*, (The Hague, The Netherlands), ACM Press, 2000.

[5] Cunha, C., Bestavros, A., and Crovella, M., *Characteristics of WWW Client-based Traces*, Boston University Computer Science Department Technical Report TR-95-010, June, 1995.

[6] Donath, J. S., and Robertson, N., "The Sociable Web," in *Proceedings of the Second International WWW Conference*, (Chicago, IL), Elsevier, 1994.

[7] Foner, L., "Yenta: A Multi-Agent, Referral Based Matchmaking System," in *Proceedings of Agents 97*, (Marina del Rey, CA USA) 1997.

[8] Gooey. Product Review Available at http://www.zdnet.com/products/stories/reviews/0,4161,2408773,00.html (Accessed Feb. 20, 2002).

[9] Grudin, J., "Groupware and Social Dynamics: Eight Challenges for Developers," *Communications of the ACM*, 37(1), 92–105, 1994.

[10] Hattori, F., Ohguro, T., Yokoo, M., Matsubara, S., and Yoshida, S., "Socialware: Multiagent Systems for Supporting Network Communities," *Communications of the ACM*, 42(3), 1999.

[11] Huberman, B., Pirolli, P., Pitkow, J., and Lukose, R., "Strong Regularities in World Wide Web Surfing," *Science*, 280(5360), 95–97, 1998.

[12] Jung, Y., and Lee, A., "Design of a Social Interaction Environment for Electronic Marketplaces," in *Proceedings of DIS 2000 - Designing Interactive Systems: Processes, Practices, Methods, Techniques*, ACM, 2000.

[13] Palfreyman, K., and Rodden, T., "A Protocol for User Awareness on the World Wide Web," in *Proceedings of CSCW 96*, (Cambridge, MA USA), ACM Press, 1996.

[14] Phelps, T., and Wilensky, R., "Robust Intra-document Locations," in *Proceedings of Ninth International World Wide Web Conference*, (Amsterdam, The Netherlands), 2000.

[15] Porter, M. F., "An Algorithm for suffix stripping," in: Spark Jones, K., and Willet, P., ed., *Readings in Information Retrieval*. San Francisco: Morgan Kaufmann, 1997.

[16] Salton, G., and Buckley, C., "Improving Retrieval Performance by Relevance Feedback," in: Spark Jones, K., and Willett, P., ed., *Readings in Information Retrieval*. San Francisco, CA: Morgan Kauffmann, 1997.

[17] Salton, G., and Buckley, C., "Term-Weighting Approaches in Automatic Text Retrieval," in: Spark Jones, K., and Willett, P., ed., *Readings in Information Retrieval*. San Francisco, CA: Morgan Kaufmann, 1997.

[18] Zipf, G., *Human Behavior and the Principle of Least-Effort*. Cambridge, MA, USA: Addison-Wesley, 1949.